Census 2000 Evaluation C.1
April 15, 2003

The Effects of the Disclosure Limitation Procedure on Census 2000 Tabular Data Products (Abridged)

FINAL REPORT

This evaluation study reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

April 15, 2003

Phil Steel
Laura Zayatz
Statistical Research Division

USCENSUSBUREAU
Helping You Make Informed Decisions

CONTENTS

EXECUTIVE SUMMARY

This report contains very limited information. It is a subset of the full Census 2000 Evaluation C.1. The full report for this evaluation is not available because it contains proprietary information. Most information in the full Census 2000 Evaluation C.1. is Census Confidential. The full evaluation cannot be removed from Census Bureau facilities and is available to Census Bureau personnel on a need-to-know basis. This abridged report is primarily descriptive and qualitative. Quantitative information can only be found in the unabridged evaluation.

Data swapping was used to protect the confidentiality of the Census 2000 tabulations. The procedure was performed on the underlying microdata, and all tabulations from the 100 percent (short form) and from the sample (long form) data were created from the swapped files. It affected pairs of households (or partnered households) where one or both of those households had a high risk of disclosure. The set of census households that were deemed as having a disclosure risk was selected from the internal census data files. These households were unique in their geographic area (block for 100 percent data and block group for sample data) based on certain characteristics. The data from these households were swapped with data from partnered households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped is not public information. The swapping procedure was performed independently for the 100 percent data and the sample data. To maintain data quality, there was a maximum percent of records that were swapped for each state for the 100 percent data and another maximum percent for the sample data.

Presumably, the higher the rate of swapping, the greater the confidentiality protection but the lower the data quality. However, the way the procedure is targeted to records with disclosure risk and the choice of variables that are controlled on and of those that are not swapped also affect the resulting levels of protection and quality. Our main goal was to see if we were able to strike the right balance between protecting confidentiality and maintaining data quality.

To answer questions on data quality, we compared tables from swapped versus unswapped data, examined the changes in cell values due to the swapping for cells of different sizes, and compared swapped and unswapped sample estimates of 100 percent data items. We also compared the effects of swapping among different geographic levels. To answer questions on data protection, we looked at how often we were able to swap households we felt had a high disclosure risk. Some calculations were performed on all 50 states. For the most detailed analysis, we used three states (Oklahoma, Massachusetts, and Mississippi) for the 100 percent data and three states (West Virginia, New Jersey, and Vermont) for the sample data.

Our key findings follow.

The data swapping procedure was checked for quality. It was conducted correctly and consistently. Minimum but necessary changes were made to the data in such a way that maximized data quality.

The disclosure limitation model used for Census 2000 is useful, and the Census Bureau should continue future research on disclosure limitation techniques. The Census Bureau should always include confidentiality protection as part of the process when planning a census.

For the 100 percent data, all records were given a chance of being swapped. The percent of records that were swapped is Census Confidential. The swapping was applied consistently in each state. Records were assigned a level of disclosure risk from 1 to 4 with 4 having the most disclosure risk. The procedure for assigning the levels of disclosure risk is Census Confidential. All level 4 records were swapped. The performance on levels 3, 2, and 1 varied from state to state and was generally better for urban states with a diverse population.

In the block level tables we examined, a small percent of cells experienced a change in value due to the swapping. The vast majority (82 percent) of cells in block-level tables are zeros and remain zeros after swapping. Of the nonzero cells, a large percentage of cells are unchanged. For tract and county tables, the average percent changes in the cell values were small. Most changes occurred in cells with small values where the disclosure risk is greatest. If a cell does change, the percent change in value depends on the original cell size. For example, a cell of size 10 might increase or decrease by 25 percent whereas a cell of size 2000 might increase or decrease by 0.5 percent.

For the sample data, all records were given some chance of being swapped. A small percent of households were swapped in each state. Again records were assigned a level of disclosure risk. Records were chosen for swapping based on their level of disclosure risk and the ability to pair records with high levels of disclosure risk. Most records deemed as having a disclosure risk were swapped.

Using variables that are common to both the 100 percent and sample data, we found that the confidence interval about the swapped sample estimate covers the true 100 percent value nearly as often as the interval about the unswapped estimate. Results were better in urban states with a diverse population.

1. BACKGROUND

This evaluation provides a discussion of measures which were used to help determine whether data swapping protected the confidentiality and preserved the data quality of Census 2000 tabulations.

1.1 Disclosure limitation

The Bureau of the Census collects decennial census data under Title 13 of the U.S. Code [Committee on Post Office and Civil Service, House of Representatives, 1991] which states that the Census Bureau shall not "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified." At the same time, the Census Bureau's mission is "to be the preeminent collector and provider of timely, relevant, and quality data about the people and economy of the United States"[1]. In order to publish as much high quality data as possible while not violating the promise of confidentiality, the Census Bureau applies disclosure limitation procedures to all data products prior to their release. The disclosure limitation procedure used to protect Census 2000 tabulations was data swapping. Note that data swapping was also used to protect the tabulations from the 100 percent data following the 1990 decennial census [Griffin et al, 1989].

1.2 Data swapping

Data swapping occurred before all tabulations from the 100 percent and from the sample data were created. It affected pairs of households (or partnered households) where one or both of those households had a high risk of disclosure. The set of census households that were deemed as having a disclosure risk was selected from the internal census data files. These households were unique in their geographic area (block for 100 percent data and block group for sample data) based on certain characteristics. We call these characteristics our targeting criteria for determining which households were at risk of disclosure. The data from these households were swapped with data from partnered households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped is not public information. The swapping procedure was performed independently for the 100 percent data and the sample data. To maintain data quality, we set a maximum percent of records that were swapped for the 100 percent data and another for the sample data. For efficiency, we tried to swap records that were at risk of disclosure with other records that were also at risk. If we could not find a partner with disclosure risk for a given household with disclosure risk, then we would resort to the set of households deemed not at risk of disclosure to search for a partner.

1.3 What this evaluation studies

Data swapping was used to protect the confidentiality of the Census 2000 tabulations. Presumably, the higher the rate of swapping, the greater the confidentiality protection but the

[1] http://landview.census.gov/contacts/www/c-mission.html

lower the data quality. However, the way the procedure is targeted to records with disclosure risk, and the choice of variables that are controlled on and of those that are not swapped also affect the resulting levels of protection and quality. Our main goal was to see if we were able to strike the right balance between protecting confidentiality and maintaining data quality.

2. METHODS

2.1 One hundred percent data

To assess protection, we used the summaries created during production. These summaries contained the counts from selection for swapping and counts from pairing the households to be swapped.

For examining data quality in this evaluation, our source files were the swapped and unswapped Hundred Percent Detail Files for three states: Oklahoma, Massachusetts, and Mississippi. These states were chosen for their contrasting performance during the swapping, due in large part to different geographical structure. We examined two tabulations; one that involved our targeting criteria, and one that was independent of the targeting criteria. These tables (several million cells), swapped and unswapped, were the basis of our evaluation.

2.1.1 Evaluating protection

All households had some chance of getting swapped. Households that had unique combinations of the characteristics in our targeting criteria were deemed at risk of disclosure and were selected and assigned a measure of disclosure risk from 1 to 4, with 4 having the most disclosure risk.

Due to the tremendous variation between states in diversity and population density, the criteria for selection behaved very differently from state to state, but generally tended to substantial overselection, that is, we selected more households than we wanted to swap. While some would argue that disclosure avoidance should be driven purely by disclosure risk, in practice, the concern over data quality led us to impose a limit on the amount of swapping that could be done within a state. Two lines of argument support this position. First, some assurance of data quality is required in order for the practice to be accepted by data users, particularly those using tabulations for setting program levels and for judicial process. Second, we felt that data quality should be as uniform as possible for all states. Hence, every state was capped at (roughly) the same percent of households being swapped, with some leeway given to achieve good performance from the pairing program. The swapping was applied consistently in each state.

Our primary measure of protection is the percentage of records of each given level of disclosure risk (1, 2, 3, or 4) that were in fact swapped. Households assigned a measure of disclosure risk of 4 were swapped with certainty.

2.1.2 Measuring data quality

The data products on which we focus for measuring data quality are tables of the type exemplified by Summary Files (SFs) 1 and 2. It is natural to break tables down to their component cells and ask the questions: How often does the published, swapped value in a cell differ from the unswapped value in that same cell and by how much? This leads to the question: "is there a relationship to our selection criteria for swapping and data quality?", i.e. how often were households with particular characteristics swapped? Various measures have been tried, the

"D" statistic [Navarro et al, 1988] , or examination of majority change; the former is somewhat abstract, the latter tailored for a specific purpose.

In tables, we can look at the data swapping procedure as introducing a type of noise into the data and view the swapped cell value as an estimate of the unswapped (true) 100 percent data value. Then for a fixed interval length, we determined the percentage of times which the unswapped values are captured by the interval when it is placed around the corresponding swapped values. We generated such intervals for the different states, different geographic levels, different variables, and different size cells. This led us to be able to make such statements as, "For Oklahoma tracts and cells in the range 116-178, 95 percent of unswapped values are within (some fixed interval length) X of the corresponding swapped values."

We also examined the average change due to the swapping in nonzero cells for different geographic levels, different variables, and different size cells for the three states.

2.2 Sample data

To evaluate protection, we again used the summary provided by the pairing program.

For examining the data quality of the sample data in this evaluation, we had three states (West Virginia, New Jersey, and Vermont) with both swapped and unswapped Sample Detail Files (SDF) available. Thus we included both urban, diverse and more rural, homogeneous states.

2.2.1 Evaluating protection

To preserve data quality, for the sample data swapping, we required that paired households matched on a larger group of variables than was used for the 100 percent data swapping. The price paid for the additional control was to raise the level of geography considerably (paired households were geographically further away), and in a few cases eliminating the ability to form pairs entirely. Our main goal was to protect the set of tract level tables in SF4. These are the largest tables published from the census at a low geographic level, somewhat less detailed than SF3 but with an additional dimension, race.

The measures of protection are the percentage of records that qualified under a particular selection criterion that were in fact swapped and the number of selected households failing to find a partner because of our requirements that partnering households match on certain characteristics.

2.2.2 Measuring data quality

We have a set of data items common to both the 100 percent and sample data. Thus we could compare the census (100 percent) tables with tables of estimates coming from both sample data sets (before and after the swapping). For any given cell we had the census (100 percent) number, the cell as it will appear in a summary file generated from the swapped sample data, and its value if we had used the unswapped sample file to create the table. We then generated [Thompson, 1991] 90 percent confidence intervals around both sample data estimates to see how often the intervals around the swapped versus unswapped estimates contained the 100 percent data value.

We examined each state separately, starting with the overall demographics and the swapping rates for different race groups. For our analysis at the state level table, we simply examined the census (100 percent) value, the swapped sample estimate, the unswapped sample estimate, and whether the 100 percent value is covered by the two confidence intervals. For county and tract data, we examined coverage rates for several different characteristics.

3. LIMITS

This section outlines the operational limits on the evaluation and limits on its distribution.

3.1 Operational limits

Because of the large amount of data and extremely large data sets, we had to limit our analysis to three states for the 100 percent data and three states for the sample data. We chose states representing a variety of race distributions and geographic sizes at different geographic levels (blocks, tracts, counties).

3.2 Census Confidential information

This report contains very limited information. It is a subset of the full Census 2000 Evaluation C.1. Most information in the full Census 2000 Evaluation C.1. is Census Confidential. The full evaluation cannot be removed from Census Bureau facilities and is available to Census Bureau personnel on a need-to-know basis.

4. RESULTS

The data swapping procedure was checked for quality. It was conducted correctly and consistently. Minimum but necessary changes were made to the data in such a way that maximized data quality.

The disclosure limitation model used for Census 2000 is useful, and the Census Bureau should continue future research on disclosure limitation techniques. The Census Bureau should always include confidentiality protection as part of the process when planning a census.

Full evaluation results are found in the Census Confidential version of this evaluation. The results below are severely limited due to confidentiality requirements.

4.1 Data protection for the 100 percent data

Households with a disclosure risk measure of 4 were swapped with certainty. The percentage of households in blocks with only one household was a strong indicator for how well overall protection goals were met, that is what percentage of households (of any type) deemed at risk we were able to swap. Low population density, with respect to census geography, increased the contribution of selected cases for all risk criteria. Where this occurred, the number of households with measures 3, 2, and 1 that were swapped would have to be smaller to fall below the maximum percent of households to be swapped. This mainly occurred in rural states. In states with urban geography, many more households with measures 3, 2, and 1 were swapped.

The confidential version of this evaluation gives explicit percentages for the rate at which households assigned the different levels of risk were swapped for all fifty states. It has a record of the percentage swapped of particular target populations (based on our targeting criteria), swap efficiency (that is the ability to pair records with disclosure risk with other records with disclosure risk), what percentage of households resided in blocks of size one, the number of unique records, the percentage of swaps within tract, and the percentage of swaps within county.

4.2 Data quality for the 100 percent data

4.2.1 Block level data

The aspect we addressed first is whether the procedure produced any global changes to the tables; specifically whether it affected the overall sparseness of the block level tables. Unrestricted data swapping could have the effect of smoothing the data, reducing its natural concentration and decreasing the empty parts of the tables.

We found no such effect for our procedure, primarily due to the forced agreement of household size. There are approximately the same number of large cell values (ten or more people) before and after swapping. There is a noticeable effect in smaller cells, which is contrary to initial expectation. Our procedure targeted households with unique characteristics, and this increased the number of zero cells (for these characteristics) because it tended to draw people from cells with values 1, 2, and 3 (presumably where the entire contribution to the cell is from one household and hence selected for high disclosure risk) and swap them into cells with larger values.

Using tables with different characteristics, we grouped table cells by value and crossed this by the absolute value of the difference between the swapped and unswapped cell values, also in ranges. The 0-0 cells, empty in both tables, were the bulk of all cells. Cells that had 0 change dominated the remainder. The majority of the action occurred in the cells with value less than 10.

4.2.2 Tract level data

For very small cells, we performed the same data analysis described above for block level data and found the same results.

For the larger cells, we viewed the data swapping procedure as introducing a type of noise into the data and viewed the swapped cell value as an estimate of the unswapped (true) 100 percent value. Then we generated an interval around the swapped value. The statistic of interest was the length of the interval required to capture 95 percent of unswapped values (see Section 2.1.2.). We generated such intervals for the three different states, different geographic levels, different variables, and different size cells. This led us to be able to make such statements as, “For Oklahoma tracts and cells in the range 116-178, 95 percent of unswapped values are within (some number) X of the swapped value.

We also examined the average change due to the swapping in nonzero cells for the three different states, different geographic levels, different variables, and different size cells. As anticipated, values representing unusual characteristics saw greater changes than others. Also, values representing characteristics used to target records with disclosure risk saw greater changes than those not used for selection.

We were satisfied that the length of the intervals and the average changes in cell values were sufficiently small.

4.2.3 County level data

Findings were consistent with the tract level data. Cells in the same size ranges had approximately the same lengths of intervals and the same average changes.

4.3 Data protection for the sample data

Swapping was applied consistently in each state. Here we made improvements in the methodology for dealing with over-selection (selecting more records than we could swap and still fall below our maximum percent of records to be swapped). Ironically, over-selection was not nearly as drastic as it was for some states in the hundred percent data. The process was less efficient, that is, more selected cases ended up partnering with unselected cases (records deemed not at risk). This was a direct consequence of the forced agreement of several characteristics in partnered households in an effort to maintain data quality. One of our protection measures was the number of households with disclosure risk for which we could not find a partner because of the forced agreement of characteristics. The number of such households was acceptably small.

The other protection measure is the percentage of records of each given level of disclosure risk that were in fact swapped. The pairs were prioritized, so that the records with the most disclosure risk were swapped first, together with the swaps that were protecting both partners. Priority levels were similar to the hundred percent but more complicated due to additional selection criteria. Most records deemed as having a disclosure risk were swapped.

4.4 Data quality for the sample data

We have a set of data items common to both the 100 percent and sample data. Thus we could compare the census tables with tables of estimates coming from sample data before and after the swapping. For any given cell we had the 100 percent value, the cell value as it will appear in a summary file generated from the swapped sample data, and its value if we had used the unswapped sample file to create the table. We then generated standard 90 percent confidence intervals around both sample data estimates to see how often the intervals around the swapped versus unswapped estimates contained the true 100 percent value.

We examined each of the three states separately, starting with the overall demographics and the swapping rates for different race groups. For our analysis of the state level table, we simply examined the census value, the sample estimate from the swapped data, the sample estimate from the unswapped data, and whether the value is covered by the two confidence intervals. For county and tract data, we examined coverage rates (how often the confidence intervals contained the 100 percent value) for several different characteristics.

4.4.1 State level estimates

The confidence interval generated around the swapped value contained the 100 percent value as often as the confidence interval around the unswapped value.

4.4.2 County level estimates

The percent of confidence intervals around the swapped values that cover the 100 percent values is slightly lower than the percent around the unswapped values. The difference in the two is larger in rural states than in urban states because, in urban states, we could find partnering households that were geographically closer.

4.4.3 Tract level estimates

Again, the percent of confidence intervals around the swapped values that cover the 100 percent values is slightly lower than the percent around the unswapped values. The difference in the two is larger for cells representing unusual characteristics where we would typically find households designated at risk of disclosure and swap them.

References

Committee on Post Office and Civil Service, House of Representatives [1991], U.S. Government Printing Office, Washington, DC, Section 9, page 4.

Griffin, R., Navarro, F., and Flores-Baez, L. [1989], "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 516-521.

Navarro, Alfredo, Flores-Baez, Linda and Thompson John H. [1988], “Results of Data Switching Simulation”, Proceedings of the American Statistical Association Meetings 1988.

Thompson, John H. [1991], “Appendix C - Accuracy of the Data for Sample Data Products”, STSD 1990 Decennial Census Memorandum Series, #W-37.